PoliMedia

Analysing Media Coverage of political debates by automatically generated links to Radio & Newspaper Items

Martijn Kleppe[1], Laura Hollink[2], Max Kemman[1], Damir Juric[3], Henri Beunders[1], Jaap Blom[4], Johan Oomen[4], Geert-Jan Houben[5]

[1] Erasmus Universiteit Rotterdam
kleppe@eshcc.eur.nl
kemman@eshcc.eur.nl
beunders@eshcc.eur.nl

[2] Vrije Universiteit Amsterdam
l.hollink@vu.nl

[3] London Brunel University
damir.juric@fer.hr

[4] Nederlands Instituut voor Beeld en Geluid
jblom@beeldengeluid.nl
joomen@beeldengeluid.nl

[5] TU Delft
g.j.p.m.houben@tudelft.nl

ABSTRACT

Students and researchers of media and communication sciences study the role of media in our society. They frequently search through media archives to manually select items that cover a certain event. When this is done for large time spans and across media-outlets, this task can however be challenging and laborious. Therefore, up until now the focus of researchers has been on manual and qualitative analyses of newspaper coverage. PoliMedia aims to stimulate and facilitate large-scale, cross-media analysis of the coverage of political events. We focus on the meetings of the Dutch parliament, and provide automatically generated links between the transcripts of those meetings, newspaper articles, including their original lay-out on the page, and radio bulletins. Via the portal at www.polimedia.nl researchers can search through the debates and find related media coverage in two media-outlets, facilitating a more efficient search process and qualitative analyses of the media coverage. Furthermore, the generated links are available via a SPARQL endpoint at data.polimedia.nl allowing quantitative analyses with complex, structured queries that are not covered by the search functionality of the portal, thus challenging the student to go across the academic borders and enter fields that previously have been neglected.

Keywords

Parliamentary debates, linking, mediatisation, linked data, media coverage, newspapers, radio

1. INTRODUCTION

Analysing media coverage of political debates across several types of media-outlets is a challenging task for academic students and researchers. Up until now, the focus of students has been on doing manual and qualitative research since newspaper articles have only been available in analogue format. Other media types such as radio bulletins have been neglected even more since these were hardly available to students. In recent years, archives of major Dutch newspapers, the transcripts of the Dutch parliament, and radio bulletins have been digitised and made available as open datasets. This contains an enormous advantage, as material can now be accessed from the Web. However, since the available data is very large, another challenge arises; it is a cumbersome and challenging task for students to analyse media items from different datasets both qualitatively as well as quantitatively. Therefore, we created automatically generated links between the transcripts of the parliament with two media-outlets: 1) newspapers in their original layout of the historical newspaper archive, and 2) radio bulletins of the Dutch National Press Agency (ANP), both located at the Dutch National Library. These links can be explored via the PoliMedia search user interface (SUI) which is currently online at www.polimedia.nl. The SUI allows students and researchers to search the debates by date and analyse the related media coverage, as well as search by name of a politician or any keyword and evaluate the debates in which the politicians appeared and how he or she was covered in the press.

An innovative approach of PoliMedia is that the coverage in the media is incorporated in its original form (figure 4), enabling analyses of both the mark-up of news articles as well as the photos in newspapers allowing further qualitative analyses of the media coverage. As a result, the big advantage of the PoliMedia system is that it allows students and researchers to make cross-media comparisons in a straightforward way both quantitatively and qualitatively. Earlier they had to manually search each archive separately, using the archives proprietary metadata, and decide whether or not a media item covers a certain (political) event. The focus of the assignments in the curriculum was therefore on qualitative analysis only. Working with the PoliMedia portal gives students and researchers a hands-on experience with a quantitative approach to their field of study. In addition, it provides

them with substantive insights into how media coverage varies over a large number of political events. We believe that this type of insight is best learned through interaction with the data, rather than, for example, literature study. With the PoliMedia approach researchers can go to one website where they will have access to all sources in a standardized format. While students and researchers before mainly used newspaper articles, the PoliMedia system allows them now to make cross-media analyses in a more efficient way. Furthermore, we made the automatically generated links available through a SPARQL endpoint at data.polimedia.nl, allowing quantitative analysis of for example the amount of links per year and decade or the number of links per political party enabling students to research the mediatisation of Dutch politics in an efficient manner.

2. RELATED WORK

The mediatisation of political debates has been the focal point of a growing field of disciplines, such as television researchers [1], communication scientists who are interested in doing discourse analysis or linguistics [2] and psychologists for researching the self-mediation of public persons [3]. However, due to the lack of available data on the mediation of the debates on radio and television, the focus up until now has been on newspapers. Since the introduction of digital sources that do include radio newsreels, television newscasts and current affairs programs, researchers should now be able to make cross media-comparisons between the different types of media-outlets. To make these large digital sources more accessible and more connected to each other, we build upon a set of guidelines and techniques to represent, link and publish data on the Linked Data Web [10] using so-called semantic web technology. In the domain of cultural heritage, the MultiMedian E-Culture project [11] has shown that through explicit representation of links between and within collections, cross-collection search becomes possible. Krouf et al. [12] demonstrate how various online sources of event information, containing both media and descriptions of events, can be merged using Linked Data. Noy et al. [13] describe how they represent and link hundreds of biomedical terminologies. In PoliMedia, we apply semantic web technology to connect various media datasets with a political event dataset. To find links between datasets that are so different in nature, we have developed a linking algorithm that includes named entity recognition and topic detection. For the latter, we have used an off-the-shelve tool called Mallet [14].

3. SYSTEM DESCRIPTION

The problem PoliMedia aims to resolve is the difficulty of searching a multitude of archives for cross-media analyses. In order to resolve this difficulty, we approached it from two perspectives; 1) the user perspective, and 2) the data perspective.

The user perspective

The targeted user groups are primarily students and researchers of *History*, *Communication and Media*, *Media Studies* and *Sociology of Culture, Media and the Arts*. However, the PoliMedia portal is valuable for a much wider range of Humanities and Social Sciences students and researchers who for example analyse the representation of politicians in the media or discussion of recurring themes. We also expect the system to be useful for several other disciplines, such as communication students who are interested in doing discourse analysis or linguistic aspects of media and political debates, psychologists researching the self-mediation of public persons, and even economists who nowadays pay more attention to the way politicians talk about the current economic crises. Furthermore, since all the links are available at data.polimedia.nl this data can also be used by students and researchers of computer science or related fields, interested in data analysis and visualization.

The development of the user interface of PoliMedia was based on a requirements study with five scholars/lecturers in history and political communication. The main use case appeared to be identifying politicians or debates of interest, and finding their representation in the media for qualitative analyses. This use case and its requirements were discussed with a UI-designer, which led to the design of a faceted search user interface (SUI) as depicted in figure 2. Facets allow the user to refine search results, they support the searcher by presenting an overview of the structure of the collection, as well as provide a transition between browsing and search strategies [6].

The SUI consists of three main levels:

1) the landing page where researchers can enter search terms (figure 1),

2) the results page (figure 2) with the search results, facets for refinements and a search bar for new queries and

3) the debate page (figure 3) which shows a complete debate and the linked media items. When clicking on a media item, the item will be opened in a new screen in its original lay-out (figure 4).

We evaluated a preliminary version of the interface by means of an eye tracking study [7]. This study showed that the faceted SUI enabled users to perform both known-item searches, as well as exploratory searches to analyse a topic over time. However, navigating the debates themselves proved to be rather difficult; as debates can be dozens of pages long, it was hard for users to gain an overview of the debate. To address this issue, the faceted search which was already available on the search results page (figure 2) was also introduced on the debate page (figure 3) in the final version of the interface.

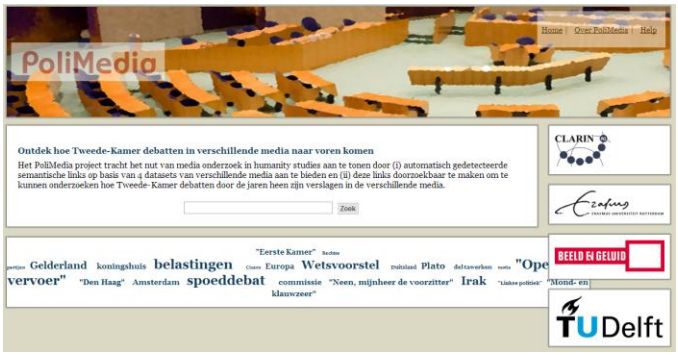

Fig. 1. *Screenshot of the PoliMedia home page*

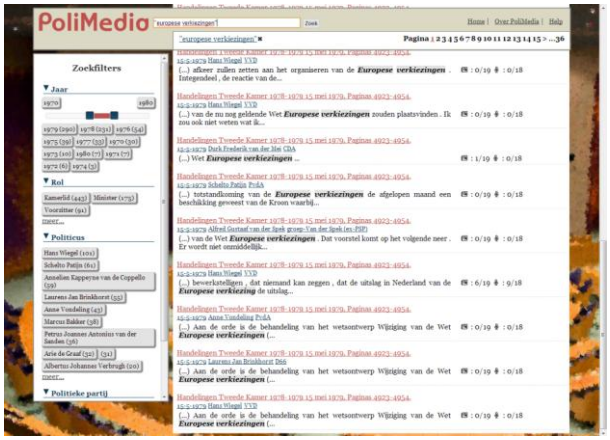

Fig. 2. *Screenshot of the PoliMedia search results page.*

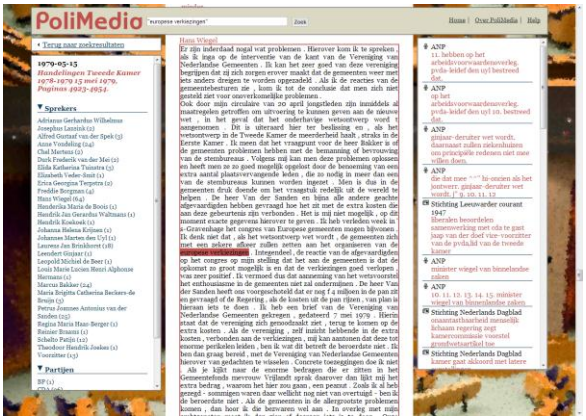

Fig. 3. *Screenshot of the PoliMedia debate page*

Fig. 4. *Screenshot of an example newspaper in original lay-out, containing an article about a parliamentary debate.*

The data perspective

In order to allow users to perform cross-media analysis in a single system, PoliMedia combines three data sources: parliamentary debates, a newspaper archive and a radio bulletin archive. The collection of Dutch parliamentary debates, the so-called *Handelingen der Staten-Generaal*, are published by the government in the form of complete transcripts of the speeches of politicians in parliamentary debates. For the period 1945-1995, the transcripts of all 9,294 debates that were held are published in unstructured TXT and PDF format at http://www.statengeneraaldigitaal.nl. The project ``War in Parliament'' has transformed them to a fine-grained XML structure [4]. We build upon War in Parliament and translate their XML to RDF. To store, query and link the debate data, we have created a semantic model in RDF which is a specialization of the more widely applicable Simple Event Model (SEM) [9]. SEM is a model that aims to represent events on the Web and explicate complicated semantic relations between people, places, actions and objects: not only who did what, when and where, but also the roles each actor played, the time during which this role is valid, and the authority according to which this role is assigned. To represent the parliamentary debates in RDF, we have created a domain specific semantic model as a specialization of SEM that enables us to express

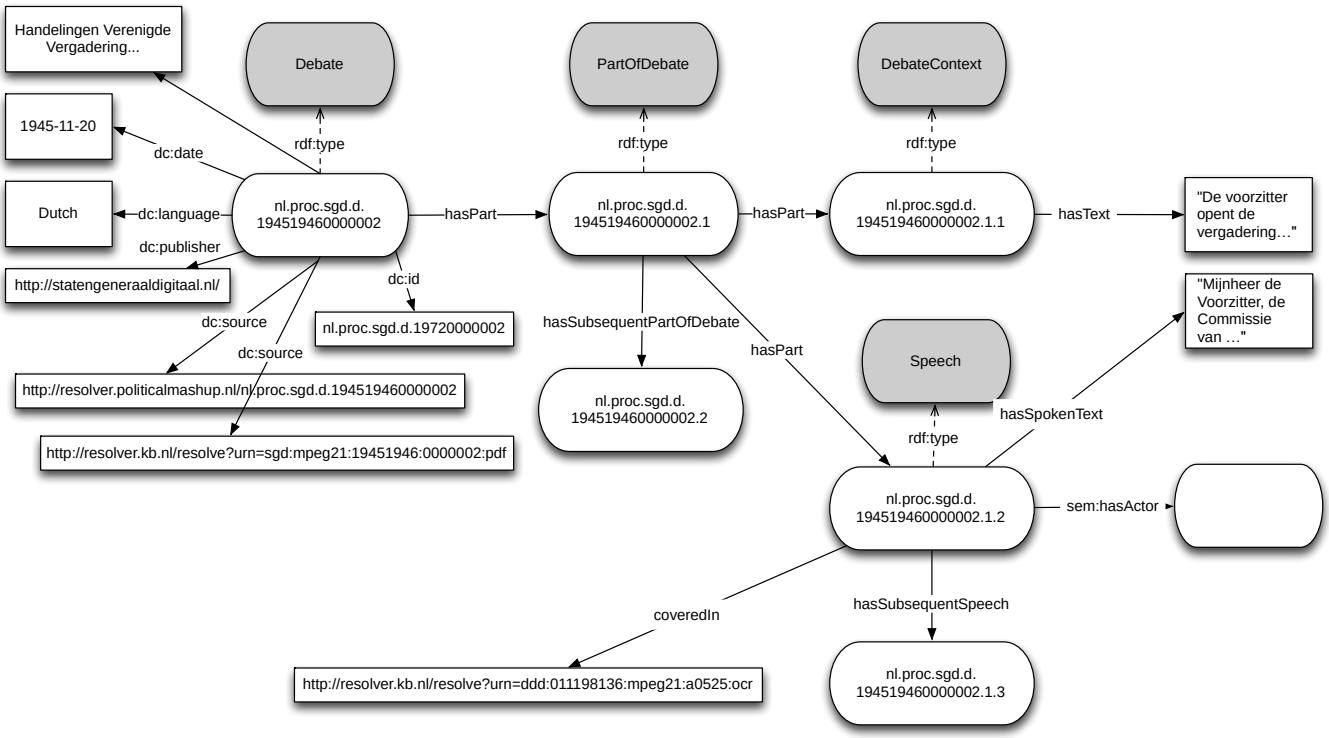

***Fig 5**: RDF model to represent parliamentary debates and links to media*

information associated with the debates such as topics, actors, debate structure, and links to media. To increase re-usability of the data, we use Dublin Core properties where appropriate, for example to denote dates, titles and publishers of debates. Figure 5 shows the RDF model. For brevity, we have left the representation of speakers (i.e. politicians and their party) out. For a detailed description of the design decisions of the model, we refer to [5] and [8].

Data Usage

The newspaper archive as well as the radio bulletin archive resides at the Dutch Royal Library. To determine links between debates and the media items in these archives, we query the full text as well as the metadata through the OAI protocol. For copyright reasons, the dataset used in the PoliMedia portal does not contain the media items themselves or their metadata; only the URIs of the items in their original archives are included. From the portal, a user can click a hyperlink to the Royal Library site to view the requested media item. At the moment, the datasets are static; they contain the debate transcripts and links to media archives of the period 1945-1995. In the future, we plan to include up-to-date data in the form of the latest debate transcripts and news articles and bulletins.

The basis of PoliMedia lies in the transcripts of the Dutch parliament from 1814-1995, containing circa 2.5 million pages of debates with speeches that have been OCR'd and thus allow for full-text search. The transcripts have been converted to structured data in XML form in previous research [4]. For each speech (i.e. a fragment from a single speaker in a debate), we extract information to represent this speech; the speaker, the date, important terms (i.e. named entities) from its content and important terms from the description of the debate in which the speech is held. This information is then combined to create a query with which we search the archives of the newspapers and radio bulletins. Media items that correspond to this query are retrieved, after which a link is created between the speech and the media item [5]. The links, as well as the parliamentary debates are represented as RDF [8]. These links are available at data.polimedia.nl as an open dataset for future researchers.

Performance

We created a stable system by using SPARQL to fetch the relevant debate data from an OWLIM repository that hosts the PoliMedia dataset. To ensure reasonable response times, the server hosting the repository has been upgraded from 8GB to 16GB of memory. Because of OWLIM's limited capabilities with respect to full-text and faceted search a separate SOLR index has been created. SOLR was chosen because of its widespread use and reputation as a high performing search index with capabilities for faceted search and many other optimization options, such as language specific options to ensure better results for Dutch.
The accuracy of our linking approach was evaluated via a manual assessment of a sample of 150 links to newspaper

articles. We found that the precision of the algorithm was good with values around 80%, with an acceptable recall of 62% [5].

Legal & Privacy

The PoliMedia portal does not involve or store any user-specific data. Since it is a web-portal, visited URLs may be stored locally by a user's own browser. Clicks on hyperlinks to media items that reside at the servers of the National library of the Netherlands may be logged by the library. The original debate data as provided by the Dutch government has a CC0 licence. The copyrights of the newspaper articles and radio bulletins are with the original publishers/broadcasters. This material may be used "for private use or a user's own study."

4. DISCUSSION

PoliMedia successfully automatically created links between the transcriptions of parliamentary debates and newspaper articles & radio bulletins, demonstrating how two very different datasets can be connected. In the near future, we intend to study the generalizability of our linking approach for linking other datasets, such as online (social) media and proceedings of other official meetings.

We already tried to link the debates with television programmes located at the Netherlands Institute for Sound and Vision but have not been able to do this. There can be several reasons for the lack of these links: the size of the available television dataset, the lack of full-text search in AV or the suitability of the linking algorithm. We expect that the metadata contained insufficient information to be linked to, while the television programs did contain coverage of the relevant debates. We hypothesize that linking to audio-visual sources requires other techniques of opening up AV archives, such as the inclusion of time-based metadata (e.g. subtitles) or the use of speech and image recognition. These techniques give more information about the content of the programs than is described in the existing metadata. We are currently working on a follow-up project of PoliMedia in which we aim to link the transcripts of the European Parliament to television programs of which the metadata has been enriched with subtitles and speech recognition to further explore the possibilities of linking to television programs.

5. CONCLUSIONS

The PoliMedia search user interface clearly shows the potential for students by linking the transcripts of political debates to different media outlets, allowing cross media analysis of both newspapers as well as radio items. However, we did not yet succeed in linking to television programmes but envision this will be possible in future research projects that can build upon the knowledge and insights we gained through the development of the PoliMedia project.

ACKNOWLEDGMENTS

The PoliMedia project was financed by CLARIN-NL and carried out by an interdisciplinary research team, consisting of both computer scientists at the TU Delft and VU Amsterdam, information scientists and historians at the Erasmus University Rotterdam and programmers at the Netherlands Institute for Sounds and Vision. We are grateful for the support of the National Library in providing the data of both the transcripts of the Dutch parliament as well as of the newspapers and radio bulletins.

REFERENCES

[1] Bignell, J., Fickers, A. (eds.) (2008). *A European Television History*. Wiley Blackwell: Malden MA / Oxford.

[2] Van Santen, R. A., Van Aelst, P., & Helfer, L. (2013). When politics becomes news: an analysis of parliamentary questions and press coverage in three West-European countries. *Acta Politica* (15 november 2013)

[3] Corner & Pels (2003) *Media and the restyling of politics* (London)

[4] Gielissen, T., & Marx, M. (2009). Exemelification of parliamentary debates. *Proceedings of the 9th Dutch-Belgian Workshop on Information Retrieval (DIR 2009)* (pp. 19–25).

[5] Juric, D., Hollink, L., & Houben, G. (2013). Discovering links between political debates and media. *The 13th International Conference on Web Engineering (ICWE'13)*. Aalborg, Denmark.

[6] Kules, B., Capra, R., Banta, M., & Sierra, T. (2009). What do exploratory searchers look at in a faceted search interface? *Proceedings of the 2009 joint international conference on Digital libraries - JCDL '09*, 313. doi:10.1145/1555400.1555452

[7] Kemman, M., Kleppe, M., & Maarseveen, J. (2013). Eye Tracking the Use of a Collapsible Facets Panel in a Search Interface. In *Research and Advanced Technologies for Digital Libraries: 17th International Conference on Theory and Practice of Digital Libraries* (pp. 401-404) Valletta: Springer Berlin Heidelberg.

[8] Juric, D., Hollink, L., & Houben, G. (2012). Bringing parliamentary debates to the Semantic Web. *DeRiVE workshop on Detection, Representation, and Exploitation of Events in the Semantic Web*.

[9] Van Hage, W. R., Malaisé, V., Segers, R., Hollink, L., & Schreiber, G. (2011). Design and use of the Simple Event Model (SEM). *Web Semantics: Science, Services and Agents on the World Wide Web*, *9*(2), 128-136.

[10] Heath, Tom, and Christian Bizer (2011). "Linked data: Evolving the web into a global data space." *Synthesis*

lectures on the semantic web: theory and technology 1.1: 1-136.

[11] Schreiber, A.T., Amin, A., Aroyo, L.M., Assem, M.F.J. van, Boer, V. de, Hardman, L., Hildebrand, M., Omelayenko, B., Ossenbruggen, J.R., Tordai, A., Wielemaker, J. & Wielinga, B.J. (2008). Semantic annotation and search of cultural-heritage collections: The MultimediaN E-Culture demonstrator. *Journal of Web Semantics*, 6(4), 243-249.

[12] Khrouf, H., and R. Troncy (2012). "EventMedia: A LOD dataset of events illustrated with media." Semantic Web journal, *Special Issue on Linked Dataset descriptions*.

[13] Noy, N. F., Shah, N. H., Whetzel, P. L., Dai, B., Dorf, M., Griffith, N., Jonquet, C. Rubin, DL., Storey, M.A., Chute, C.G., & Musen, M. A. (2009). BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic acids research*, 37(suppl 2), W170-W173.

[14] McCallum, Andrew Kachites (2002) *MALLET: A Machine Learning for Language Toolkit*, http://mallet.cs.umass.edu.